CVPR
#4469

Rethinking Imbalance in Radiology Report Generation: A Curriculum Learning Perspective

Anonymous CVPR submission

Paper ID 4469

Abstract

*Radiology report generation (RRG) is of great significance in alleviating the workload of radiologists and improving clinical diagnosis efficiency, which has attracted widespread attention in the academic community. Due to the inherent nature of medical data, datasets used for training the RRG model commonly suffer from imbalance issues. Most existing studies attribute this imbalance to the varying number of samples across classes (i.e., traditional class imbalance). However, our research reveals that the impact of such traditional class imbalance on RRG is limited, given the complexity of the task. Inspired by curriculum learning (CL), we propose to quantify data imbalance in RRG through learning difficulty and introduce a **R**adiology **R**eport **G**eneration **C**urriculum **L**earning (RRGCL) framework. Additionally, a novel dynamic data scheduling strategy is designed to help models better adapt to this defined imbalance during training. Notably, our method is model-agnostic and plug-and-play, making it potentially applicable to a wide range of RRG models. Extensive experiments demonstrate that our method further enhances the performance of several state-of-the-art RRG methods. To the best of our knowledge, our method is orthogonal to most existing RRG methods, offering a new perspective for RRG research and promising to advance the further development of intelligent medicine. The code will be made available upon acceptance.*

1. Introduction

In modern medicine, various medical images have become important tools for clinical decision-making and treatment planning. A medical radiology report is an objective and detailed description of significant findings in given medical imaging data. Writing radiology report is a highly specialized task, which must be completed by well-trained radiologists with rich professional knowledge and clinical experience. Moreover, with the widespread use of medical imaging, the workload of radiologists has also increased [4]. This is bad news, because the excessive workload tends to increase the risk of diagnostic errors [23]. In order to alleviate radiologists' work pressure, improve the accuracy of radiology reports and assist clinical diagnosis, automatic

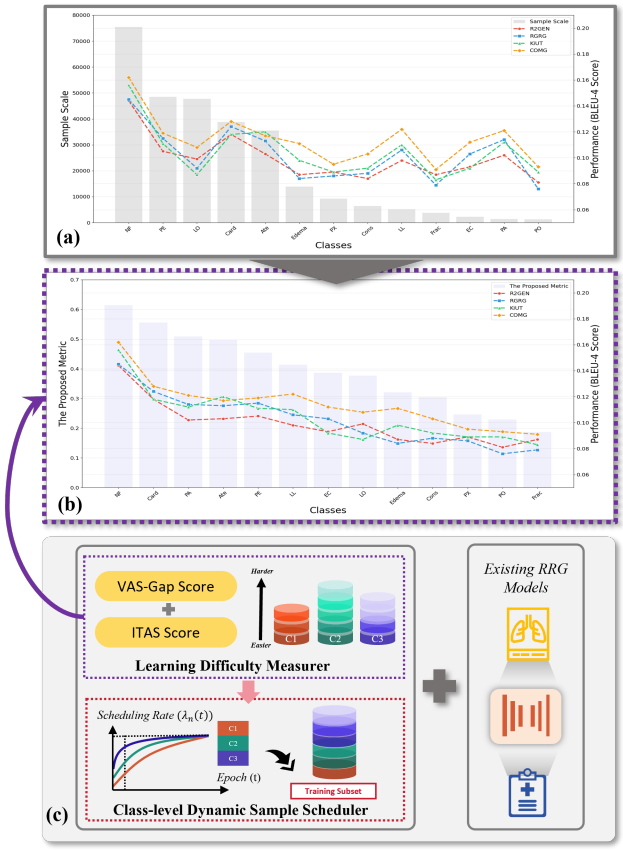

Figure 1. (a) and (b) respectively illustrate the relationship between the imbalance in RRG data measured by traditional sample scale and our proposed metric and the quality of reports generated for each classes. The comparison reveals that our proposed metric demonstrates a better proportional relationship, indicating its higher efficiency. (c) provides a glimpse of our proposed RRGCL.

CVPR #4469

radiology report generation (RRG) has received much attention [36, 43, 44, 51].

Traditional class imbalance is a prevalent issue in various medical datasets. Due to factors such as disease incidence rates and diagnosis difficulty, the data scale of different classes in medical datasets often vary significantly, making it challenging for the model to learn the features of minority classes. However, it is noteworthy that unlike general classification tasks, RRG is more complex because its labels are long texts containing extensive information, and although the classes are fixed, the content of the reports varies. Consequently, we are prompted to raise two previously overlooked questions: **Q1:***What are the impacts of traditional class imbalance on RRG models?* **Q2:** *Is there a better metric to more effectively measure the imbalance in RRG data?*

To address **Q1**, we selected four classical and state-of-the-art (SOTA) RRG methods [7, 12, 16, 47] and conducted experiments on the currently largest publicly available RRG dataset, MIMIC-CXR [21, 22]. We divided the test set into 13 classes based on the official tags and evaluated the quality of the generated reports (BLEU-4 score) for each class. We arranged the classes in descending order of their sample scale (gray bars) and observed the relationship with the quality of the generated reports, as shown in Figure 1(a). Intuitively, according to the traditional class imbalance assumption, the fewer training samples a class has, the worse its corresponding generated reports will be. In other words, the two should be almost proportional. However, as shown in the figure, this assumption does not hold significantly for the RRG task.

Based on the above finding, we propose a method of **R**adiology **R**eport **G**eneration by **C**urriculum **L**earning (RRGCL) to address **Q2**. Inspired by curriculum learning (CL) [3], we conceive a novel learning difficulty metric to reevaluate imbalance in RRG data. This metric comprises two components: (1) visual neighborhood semantic gap (VNS-Gap); (2) image-text cross-modal alignment score (ITAS). We rearrange the classes in descending order according to the proposed metric and again observe their relationship with the quality of the generated reports, with the results shown in Figure 1(b). As evidenced in the figure, the proposed metric demonstrates a stronger positive correlation with report generation quality, indicating its superior efficacy in measuring imbalance within RRG data. Furthermore, we design a novel sample scheduling function based on the proposed imbalance metric. It dynamically adjusts the pace of introducing data of different difficulties into training subset, thereby helping the model gradually adapt to the inherent imbalance in RRG data. Extensive experiments demonstrate that RRGCL can effectively help existing RRG models further improve the quality of generated reports. Overall, the main contributions of this paper are summarized as follows:

- We discover that the traditional class imbalance based on sample quantity has limitations in RRG data, and propose a novel learning difficulty metric that can more effectively measure imbalance in RRG data.
- To address the redefined imbalance issue, we propose a simple yet efficient sample scheduling strategy for RRG models.
- Our proposed RRGCL method is a model-agnostic and plug-and-play training strategy. Extensive experiments demonstrate that existing RRG methods achieve performance improvements after applying RRGCL.

2. Related Work

2.1. Radiology Report Generation

Automatic radiology report generation (RRG) originates from image captioning task, but it is more complex and challenging. Due to pressing clinical needs, research in RRG has been steadily increasing in recent years [44]. Current RGG studies can be broadly categorized into three main classes: CNN-RNN-based method, Transformer-based method and retrieval-based method.

CNN-RNN-based models [13, 35, 38, 47, 52, 56, 61] represent the most classic encoder-decoder frameworks for RRG. In this paradigm, CNN network or its variants (e.g., ResNet, DenseNet) are typically employed to extract visual features, which are then decoded by RNN network or its variants (e.g., LSTM, GRU) to to generate the report word by word. Such RRG methods are adapted from early image captioning tasks, laid a crucial foundation for subsequent research, and continue to be influential today. However, they often exhibit limitations in modeling long-range dependencies and tend to overlook subtle lesions in images. The introduction of Transformer [48] effectively mitigates this issue. Transformer-based methods [5, 7, 8, 12, 16, 24, 27, 31, 34, 37, 49, 53] represent the dominant paradigm in current RRG research. These approaches typically employ a Transformer as the decoder to generate long-text reports, while the image encoder is usually a visual Transformer (ViT) [9] or a hybrid CNN-Transformer architecture for visual feature extraction. Thanks to the attention mechanism, such methods are often better at modeling long-range dependencies, thereby generating higher-quality and more human-like reports. Due to similar patterns among radiology reports, some retrieval-based methods [10, 19, 46, 58, 62] have been introduced. These models first retrieve similar reports or sentences from a template or report corpus, and then use this as reference knowledge to generate the final report. These methods are heavily dependent on large-scale, specialized databases and exhibit limited generalization capability.

Our proposed RRGCL method falls outside the afore-

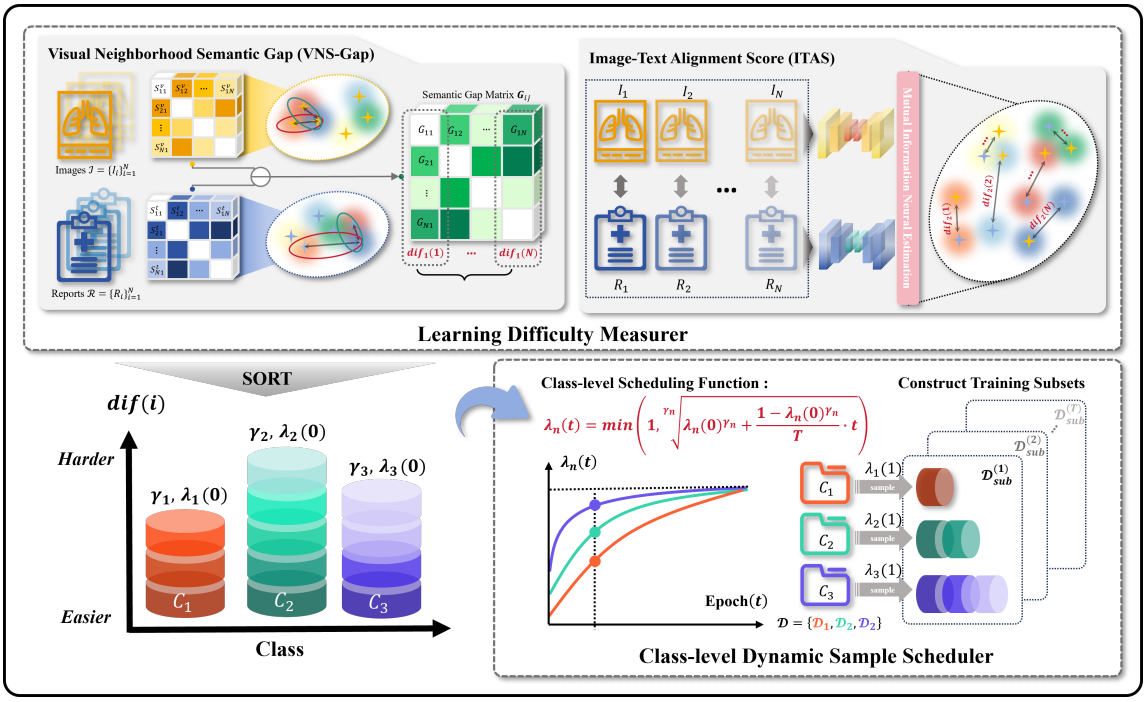

Figure 2. The overall framework of the proposed RRGCL primarily consists of two components: a Learning Difficulty Measurer and a Class-level Dynamic Sample Scheduler. The measurer evaluates the learning difficulty $dif(i)$ of each sample by calculating VNS-Gap $dif_1(i)$ and ITAS $dif_2(i)$. Based on $dif(i)$, all training samples are sorted from easy to hard within each class. The scheduler dynamically samples from each class in sequence according to our designed class-level scheduling function $\lambda_n(t)$. Simpler classes are incorporated into the training subset at a faster rate during the early training stages, while more difficult classes are introduced at a more gradual pace. Class difficulty scores reflects the underlying imbalance in RRG data from a novel perspective, and the proposed scheduling strategy effectively addresses this imbalance.

mentioned classes, because it cannot generate reports by itself. Instead, it is a model-agnostic training strategy that can be applied to various existing RRG models to further enhance their performance.

2.2. Curriculum Learning

The core idea of curriculum learning (CL) [3] is to mimic the human cognitive process of progressing from easy to difficult. In the early stages of training, the model primarily encounters the simplest, cleanest, and most representative samples in the dataset, and then gradually incorporates more complex data. This strategy provides the model with a better initial optimization path, leading to more stable and efficient training [45, 50].

CL has been proven effective in various downstream tasks. For example, in the field of computer vision (CV), it has been applied successfully to medical image classification [20, 26], face recognition [15], object detection [59], etc. In the field of natural language processing (NLP), it has shown effectiveness in machine translation [11], text emotion recognition [25, 57], text generation [6], etc. Additionally, it has also been applied to some cross-modal tasks like image caption [55] and video caption [28]. To the best of our knowledge, CMCL [32] is currently the only CL-based method for RRG. CMCL defines two training difficulty evaluation metrics for images and text respectively. During training, it sorts all training data into four batches based on these metrics and dynamically selects the most appropriate batch for training according to the model's current learning competence. Our proposed method does not evaluate images and text in isolation. Instead, it calculates the learning difficulty of each sample by leveraging inter-sample divergence and image-text relationship.

3. Method

RRGCL mainly consists of two parts: a learning difficulty measurer and a dynamic sample scheduler. The learning difficulty measurer quantitatively evaluates how challenging it is to learn from a class, thereby revealing the underlying imbalance in the RRG data. The dynamic sample scheduler determines how to progressively increase the learning difficulty during training, which we utilize to mitigate the imbalance issue. The overall framework of RRGCL is shown in Figure 2.

CVPR #4469

3.1. Definition

Let $\mathcal{D} = \{I_i, R_i, y_i\}_{i=1}^{N}$ denote the training dataset including N samples, where $I_n \in \mathcal{I}$ represents the i-th radiology image and $R_i \in \mathcal{R}$ represents the corresponding i-th report, and y_i is the label including c classes. The objective of RRG is to learn a mapping function $\mathcal{G}_\theta : \mathcal{I} \rightarrow \mathcal{R}$ that generates reports from input images.

3.2. Learning Difficulty Measurer

3.2.1. What Does Difficult RRG Data Look Like?

Before elaborating on our method, it is necessary to consider this question: *What kind of RRG data might be difficult for the model to learn?* We answer this question from two perspectives. (1) Visual differences between medical images are often subtle, yet the descriptions of these minor variations in reports can be substantially distinct. For example, a slight increase in the density of a region might be described as new-onset inflammation in the report. Therefore, if an RRG sample exhibits visual similarity to other cases but is associated with a significantly divergent report, it is likely difficult to learn, as such samples create confusion for the model. (2) RRG is fundamentally a cross-modal transfer task from visual to textual data, where the degree of inter-modal alignment directly influences performance. Thus, samples with stronger image-text alignment are generally easier to learn. Based on these considerations, we design two specialized metrics to quantify the learning difficulty of RRG data: Visual Neighborhood Semantic Gap (VNS-Gap) and Image-Text Alignment Score (ITAS).

3.2.2. Visual Neighborhood Semantic Gap

The visual neighborhood semantic gap (VNS-Gap) quantifies the learning difficulty by measuring the discrepancy between visual similarity and semantic similarity across samples. Let $\phi : \mathcal{I} \rightarrow \mathbb{R}^d$ be a pre-trained visual encoder, and the visual similarity between two images I_i and I_j is computed as follows:

$$\mathbf{S}_{ij}^{v} = \frac{\phi(I_i)^\top \phi(I_j)}{\|\phi(I_i)\|_2 \cdot \|\phi(I_j)\|_2} \tag{1}$$

where $\mathbf{S}_{ij}^{v} \in \mathbb{R}^{N \times N}$ represents the visual similarity matrix. Similarly, let $\psi : \mathcal{R} \rightarrow \mathbb{R}^d$ be a a pre-trained textual encoder, and the textual similarity between two reports R_i and R_j is computed as follows:

$$\mathbf{S}_{ij}^{t} = \frac{\psi(R_i)^\top \psi(R_j)}{\|\psi(R_i)\|_2 \cdot \|\psi(R_j)\|_2} \tag{2}$$

where $\mathbf{S}_{ij}^{t} \in \mathbb{R}^{N \times N}$ represents the textual similarity matrix. To capture the absolute discrepancy between visual and textual similarities, a semantic gap matrix $\mathbf{G}_{ij} \in \mathbb{R}^{N \times N}$ is calculated as follows:

$$\mathbf{G}_{ij} = |\mathbf{S}_{ij}^{v} - \mathbf{S}_{ij}^{t}| \tag{3}$$

Finally, the VNS-Gap of the i-th sample can be easily measured as follows:

$$dif_1(i) = \sum_{j=1}^{N} \mathbf{G}_{ij} \tag{4}$$

3.2.3. Image-Text Alignment Score

Mutual Information (MI) is an efficient statistical measure for calculating the image-text alignment score (ITAS) between radiology images and corresponding reports. The MI between image I and report R is defined as follows:

$$M(I; R) = D_{KL}(\mathbb{P}_{IR} \| \mathbb{P}_I \otimes \mathbb{P}_R) \tag{5}$$

where $\mathbb{P}_{IR}$ denotes the joint distribution of image-text pairs, $\mathbb{P}_I \otimes \mathbb{P}_R$ represents the product of marginal distributions, and D_{KL} is the Kullback-Leibler (KL) divergence. However, estimating MI between high-dimensional continuous variables with limited data is quite challenging [29], so we employ Mutual Information Neural Estimation (MINE) [18] to circumvent the issue of the dimensionality disaster. MINE leverages the Donsker-Varadhan (DV) dual representation of the KL-divergence, which provides a lower bound for MI as follows:

$$M(I; R) \geq \sup_{T \in \mathcal{F}} \left[\mathbb{E}_{\mathbb{P}_{IR}}\left[T(I, R)\right] - \log\left(\mathbb{E}_{\mathbb{P}_I \otimes \mathbb{P}_R}\left[e^{T(I,R)}\right]\right)\right] \tag{6}$$

where $T : \mathcal{I} \times \mathcal{R} \rightarrow \mathbb{R}$ is a function in the family $\mathcal{F}$ of bounded measurable functions. We parameterize the function T using a neural network T_θ with parameters θ, transforming the estimation into an optimization problem as follows:

$$\hat{M}_\theta(I; R) = \mathbb{E}_{\mathbb{P}_{IR}}\left[T_\theta(I, R)\right] - \log\left(\mathbb{E}_{\mathbb{P}_I \otimes \mathbb{P}_R}\left[e^{T_\theta(I,R)}\right]\right) \tag{7}$$

The optimal parameters are obtained by solving:

$$\theta^\star = arg\max_{\theta} \hat{M}_\theta(I; R) \tag{8}$$

After training, we obtain an image-text alignment scorer $T_{\theta^\star}$, which evaluates i-th sample (I_i, R_i) as follows:

$$dif_2(i) = T_{\theta^\star}(I_i, R_i) \tag{9}$$

3.2.4. Overall Learning Difficulty

The overall learning difficulty score for i-th sample is obtained by combining both VNS-Gap and ITAS after normalization:

$$dif(i) = \alpha \cdot \frac{dif_1(i) - \mu_1}{\sigma_1} + \beta \cdot \frac{dif_2(i) - \mu_2}{\sigma_2} \tag{10}$$

where (μ_1, σ_1) and (μ_2, σ_2) are the mean and standard deviation of $dif_1(i)_{i=1}^{N}$ and $dif_2(i)_{i=1}^{N}$, respectively. $\alpha, \beta \in \mathbb{R}^+$ are weighting coefficients, which will be discussed in Section 4.6.

CVPR #4469

3.3. Class-level Dynamic Sample Scheduler

Most sample scheduler is to sort all training data from easy to hard based on learning difficulty and then schedule the data sequentially according to a certain rule. This paradigm may result in some classes with fewer samples rarely or never appearing in the newly added training subsets during the later training stages, which is likely to exacerbate data imbalance. To end this, we propose a class-level dynamic sample scheduling strategy, as shown in Algorithm 1.

Algorithm 1 Sample Scheduling Strategy of RRGCL

Input: The training set $\mathcal{D}$, the number of training epochs T, the training difficulty $dif(i)$ for each sample.
Output: A RRG model $\mathcal{G}_\theta$.

```
Divide D into c-class subsets, resulting in D = {D_n}_{n=1}^c;
Sort samples in D_n based on the sample difficulty and calculate class difficulty dif_C^(n) based on dif(i);
Calculate "scheduling exponential" γ_n and initial scheduling rate λ_n(0);
Calculate scheduling function λ_n^(t) by Eq.11;
D_sub^(t) = ∅;
for t = 0 to T do
    for n = 1 to c do
        Extract the top λ_n(t) proportion of training samples from each class;
        D_sub^(t) ← D_sub^(t+1);
    end for
    Train G_θ on D_sub^(t);
end for
```

3.3.1. Designing Class-level Scheduling Function

We divide the training set $\mathcal{D}$ into c subsets according to the class label y_i, and then arrange the samples within each subset in ascending order of learning difficulty $dif(i)$, resulting in $\mathcal{D} = \{\mathcal{D}_n\}_{n=1}^c$. A class-level learning difficulty metric $dif_C^{(n)}$ is introduced to measure the difficulty of each class by averaging the difficulties of all samples within that class.

The root function [42] has been proven to be an effective function for sample scheduling. Our proposed scheduling strategy assigns each class a root function $\lambda_n(t)$ based on its class-level difficulty, thereby controlling the sampling rate from different classes. The specific is as follows:

$$\lambda_n(t) = min\left(1,\ \sqrt[\gamma_n]{\lambda_n(0)^{\gamma_n} + \frac{1-\lambda_n(0)^{\gamma_n}}{T}\cdot t}\right) \quad (11)$$

where T represents the total number of training epochs, the variable $t \in [1, T]$ represents the current training epoch, and the value range of $\lambda_n(t)$ is $(0, 1]$. $\gamma_n \in [1, c+1]$ is a "scheduling exponential" we defined, which is a function of the class-level difficulty $dif_C^{(n)}$, as shown below:

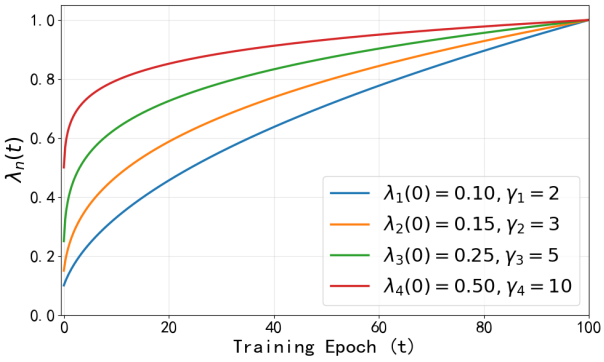

Figure 3. An example of scheduling function $\lambda_n(t)$ over training epochs for different γ_n and $\lambda_n(0)$, where $T = 100$.

$$\gamma_n\left(dif_C^{(n)}\right) = 1 + \left(1 - dif_C^{(n)}\right)\cdot c \quad (12)$$

This is a simple yet efficient method that ensures simpler classes are assigned larger "scheduling exponential". $\lambda_n(0)$ is the initial scheduling rate for the n-th class can be set according to γ_n as follows:

$$\lambda_n(0) = \frac{\gamma_n}{\sum_{i=1}^{c}\gamma_n} \quad (13)$$

3.3.2. Constructing Training Subsets

For each training epoch t, the scheduler constructs a training subset $\mathcal{D}_{sub}^{(t)}$ by sequentially selecting the top $\lambda_n(t)$ proportion of samples from each class. To intuitively illustrate the scheduling strategy we designed, a straightforward example with manually configured parameters is provided, as shown in Figure 3. It can be observed that simpler classes (γ_n is bigger) are scheduled in a more "aggressive" manner, with a faster scheduling rate in the early training stages, while difficult classes (γ_n is smaller) are scheduled more gradually, progressing at a steady rate throughout the training process. At the T-th epoch, all samples join the training. By constructing training subsets in this manner, the model is exposed to simpler samples in the early training stages. In the later stages, the newly added samples still encompass data from all classes. Furthermore, due to the characteristics of the root function, the increase in difficulty throughout the training process is more gradual, ensuring training stability.

4. Experiment

4.1. Datasets

All experiments are conducted on the MIMIC-CXR dataset [21, 22], which is currently the largest publicly available RRG dataset, containing 377,110 chest X-ray images and 227,835 radiology reports from 63,478 patients. MIMIC-CXR provides 14 uncertainty labels extracted by CheXpert

CVPR #4469

Table 1. The performance improvement of SOTA RRG methods by integrating RRGCL on the MIMIC-CXR dataset.

Method	NLG Metrics					CE Metrics			
	BL-1↑	BL-4↑	MTR↑	RG-L↑	Avg.Δ	P↑	R↑	F1↑	Avg.Δ
R^2GEN[7] (EMNLP'2020)	0.353	0.103	0.142	0.277	-	0.333	0.273	0.276	-
R^2GEN + RRGCL	0.361	0.108	0.152	0.284	**+3.429%**	0.341	0.280	0.282	**+2.380%**
KiUT[16] (CVPR'2023)	0.393	0.113	0.160	0.285	-	0.371	0.318	0.321	-
KiUT + RRGCL	0.402	0.117	0.169	0.293	**+3.260%**	0.382	0.327	0.329	**+2.772%**
RGRG[47] (CVPR'2023)	0.373	0.126	0.168	0.264	-	0.495	0.475	0.447	-
RGRG + RRGCL	0.388	0.129	0.177	0.278	**+4.404%**	0.512	0.489	0.459	**+4.035%**
COMG[12] (WACV'2024)	0.363	0.124	0.128	0.290	-	0.424	0.291	0.345	-
COMG + RRGCL	0.373	0.126	0.141	0.298	**+3.646%**	0.437	0.313	0.356	**+4.340%**
EKAGen[5] (CVPR'2024)	0.419	0.119	0.157	0.264	-	0.517	0.483	0.499	-
EKAGen + RRGCL	0.427	0.122	0.163	0.274	**+2.815%**	0.522	0.498	0.513	**+2.268%**
DACG[24] (MedIA'2025)	0.398	0.117	0.162	0.290	-	0.422	0.405	0.389	-
DACG + RRGCL	0.409	0.120	0.166	0.296	**+2.482%**	0.436	0.418	0.403	**+3.372%**

Table 2. Ablation studies on the learning difficulty measurer and sample scheduler conducted on the MIMIC-CXR dataset. The best results are bolded.

Method	Difficulty Measurer		Scheduler	NLG Metrics				CE Metrics		
	VNS-Gap	ITAS		BL-1	BL-4↑	MTR↑	RG-L↑	P↑	R↑	F1↑
Baseline (RGRG [47])	-	-	-	0.373	0.126	0.168	0.264	0.495	0.475	0.447
(a)	✓	-	Ours	0.381	0.127	0.173	0.272	0.508	0.482	0.454
(b)	-	✓	Ours	0.378	0.127	0.171	0.269	0.504	0.480	0.451
(c)	✓	✓	Baby Step	0.375	0.126	0.170	0.267	0.499	0.478	0.450
(d)	✓	✓	Linear	0.376	0.126	0.172	0.270	0.501	0.479	0.452
(e)	✓	✓	One Pass	0.375	0.126	0.169	0.267	0.500	0.478	0.449
(f)	✓	✓	Root	0.382	0.128	0.174	0.274	0.508	0.485	0.455
(g)	✓	✓	Geom	0.379	0.127	0.171	0.272	0.504	0.481	0.452
RGRG + RRGCL	✓	✓	Ours	**0.388**	**0.129**	**0.177**	**0.278**	**0.512**	**0.489**	**0.459**

[17]. We adopt 13 keywords excluding "Support Devices" and treat the keywords labeled as "1" (positive) as the labels for the corresponding samples. The construction of RRGCL and the training of the RRG models are independent. During the training phase of the learning difficulty measurer, all images are resized to 256 × 256, and the "Findings" and "Impressions" sections in the reports are stitched together. During the training phase of the RRG model, all data preprocessing follows the combined RRG methods.

4.2. Evaluation Metrics

Natural language generation (NLG) metrics are commonly used standards for evaluating the quality of generated reports, including BLEU-n (BL-1, BL-4) [39], METEOR (MTR) [2] and ROUGE-L (RG-L) [30]. Some studies [33, 60] point out that NLG metrics may not be well-suited for evaluating radiology report generation model. Therefore, to more accurately evaluate the proposed method, we used several clinical efficacy (CE) metrics, including precision (P), recall (R) and F1-score (F1).

4.3. Implementation Details

For the VNS-Gap measuring network, we use a ResNet101 [14] fine-tuned on CheXpert [17] to extract visual features. For the ITAS computation network, following Liao et al. [29], we employ a pre-trained 5-layer ResNet [14] to extract visual features. All textual features are extracted using the pre-trained Clinical BERT [1]. The hyperparameters α and β in Eq.10 are set to 0.6 and 0.4, respectively. During the training of the RRG models, the configuration of training parameters follows the original settings of the combined RRG methods. All training processes are performed on an NVIDIA GeForce RTX A6000 GPU.

4.4. Improvements to SOTA RRG Models

To evaluate how RRGCL helps improve the performance of RRG models, we embed RRGCL into several SOTA RRG methods including R^2GEN[7], KiUT [16], RGRG

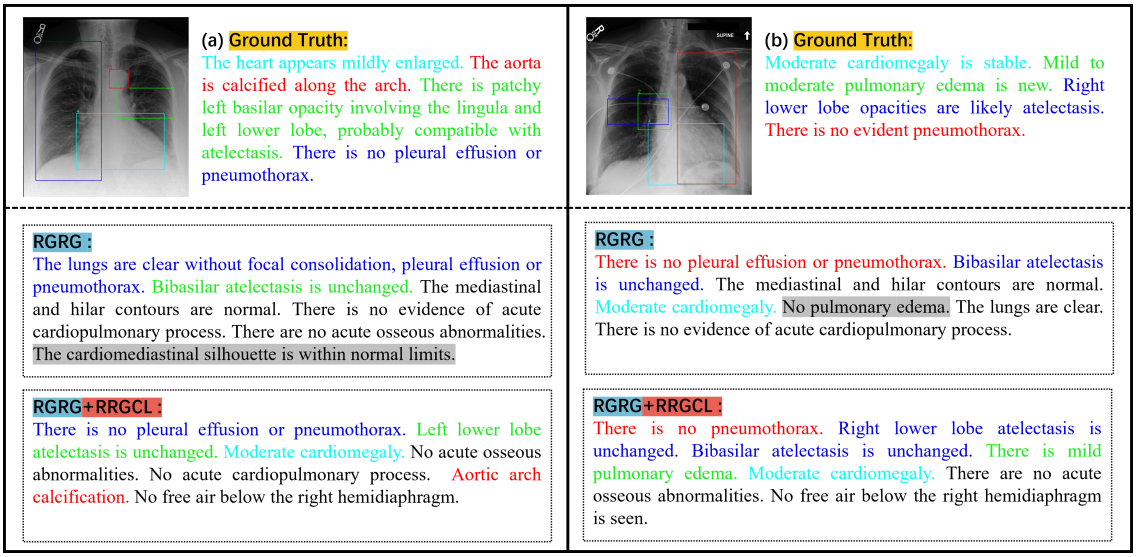

Figure 4. Two examples of reports generated by RGRG [47] and RGRG + RRGCL. The bounding boxes are provided by the ImaGenome dataset [54], and they correspond to the descriptions in the reports based on color. Sentences with gray shading represent incorrect descriptions.

Table 3. Hyperparameter sensitivity experiments on α and β based on RGRG[47]. The best results are bolded.

α	β	BL-1↑	BL-4↑	F1↑
0.2	0.8	0.382	0.128	0.451
0.4	0.6	0.385	0.128	0.454
0.6	0.4	**0.388**	**0.129**	**0.459**
0.8	0.2	0.386	0.129	0.457

[47], COMG [12], EKAGen [5] and DACG [24], and observe the quality of the generated reports. The experimental results are shown in Table 1. The results indicate that when combined with RRGCL, the performance of RRG models demonstrates significant improvements in both NLG and CE metrics. This proves that our proposed method can effectively enhance existing works and has the potential to be applied to more RRG models.

4.5. Ablation Study

To validate the effectiveness of the proposed learning difficulty measurer and sample scheduler, we conduct ablation experiments. We select RGRG [47], which has the most significant performance improvement in Table 1 as the baseline, and the results are shown in Table 2. Among them, Baby Step [3], Linear [3], One Pass [3], Root [41] and Geometric Progression [40] are scheduling strategies proposed in previous works. The experimental results of methods (a) and (b) demonstrate that our proposed VNS-Gap and ITAS can effectively characterize the learning difficulty of RRG data, with VNS-Gap being relatively more efficient. Compared to the baseline, methods (c)-(g) all achieve some improvements, while the use of our proposed scheduler yield the best performance. It is worth noting that our proposed scheduling function further allocates class-level scheduling functions based on the Root function. Compared to method (f), our method achieve improved performance because the proposed scheduling strategy not only ensures a smooth increase in learning difficulty but also maintains a balanced distribution of various class in each training epoch.

4.6. Hyperparameter Sensitivity Experiment

To determine the optimal settings for the hyperparameters α and β in Eq.10, we conduct a hyperparameter sensitivity experiment based on RGRG [47]. The results are shown in Table 3. It can be observed that the model achieves the best performance when α and β are set to 0.6 and 0.4, respectively.

4.7. Qualitative Analysis

To qualitatively analyze the effectiveness of RRGCL, we randomly selected two test samples and compared the reports generated by RGRG [47] and RGRG + RRGCL, the results are shown in Figure 4. In example (a), RGRG fails to detect "aortic calcification" and incorrectly considers the heart size to be normal. However, after introducing RRGCL, the generated report describes both. In the image of example (b), there is a region of "pulmonary edema", and the report generated by RGRG + RRGCL describes it, while RGRG provides an incorrect description. The qualitative experimental results visually demonstrate the effect of

RRGCL in enhancing the performance of the RRG model.

5. Conclusion

In this work, we focus on a previously overlooked issue in radiology report generation (RRG): the impact of data imbalance in RRG training on model performance. We discover that traditional class imbalance based on sample quantity fails to adequately characterize the imbalance in RRG data. Therefore, from a curriculum learning perspective, we redefine the imbalance in RRG data by learning difficulty. Our proposed measurer calculates the sample learning difficulty from the perspectives of visual neighborhood semantic gap and image-text alignment scores. To address the imbalance we proposed, we further introduce a class-level dynamic sample scheduler. Experimental results demonstrate that our proposed method can effectively enhance the performance of existing RRG models, offering a new direction for RRG studies. In the future, with the release of labeled RRG datasets, we will validate our method on more datasets.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78, 2019. 6

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2, 3, 7

[4] RJM Bruls and RM Kwee. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into imaging*, 11:1–7, 2020. 1

[5] Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204, 2024. 2, 6, 7

[6] Rohan Chaudhury, Maria Teleki, Xiangjue Dong, and James Caverlee. Dacl: Disfluency augmented curriculum learning for fluent text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4311–4321, 2024. 3

[7] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020. 2, 6

[8] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q O'Neil. Controllable chest x-ray report generation from longitudinal representations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. 2

[11] Xiang Geng, Yu Zhang, Jiahuan Li, Shujian Huang, Hao Yang, Shimin Tao, Yimeng Chen, Ning Xie, and Jiajun Chen. Denoising pre-training for machine translation quality estimation with curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12827–12835, 2023. 3

[12] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 7995–8004, 2024. 2, 6, 7

[13] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Lin, and Weidong Cai. Orid: Organ-regional information driven framework for radiology report generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 378–387. IEEE, 2025. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 3

[16] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2023. 2, 6

[17] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 6

[18] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv e-prints*, pages arXiv–1801, 2018. 4

[19] Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. Multimodal

image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*, pages 978–990. PMLR, 2024. 2

[20] Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A González Ballester, and Gemma Piella. Medical-based deep curriculum learning for improved fracture classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 694–702. Springer, 2019. 3

[21] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 2, 5

[22] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 2, 5

[23] Ömer Kasalak, Haider Alnahwi, Romy Toxopeus, Jan P Pennings, Derya Yakar, and Thomas C Kwee. Work overload and diagnostic errors in radiology. *European Journal of Radiology*, 167:111032, 2023. 1

[24] Wangyu Lang, Zhi Liu, and Yijia Zhang. Dacg: Dual attention and context guidance model for radiology report generation. *Medical Image Analysis*, 99:103377, 2025. 2, 6, 7

[25] Nana Lange, Flavius Frasincar, and Maria Mihaela Truşcă. Curriculum learning for a hybrid approach for aspect-based sentiment analysis. *Expert Systems with Applications*, page 129669, 2025. 3

[26] Chaoyi Li, Meng Li, Can Peng, and Brian C Lovell. Dynamic curriculum learning via in-domain uncertainty for medical image classification. In *International conference on medical image computing and computer-assisted intervention*, pages 747–757. Springer, 2023. 3

[27] Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsaddik, and Xiaojun Chang. Contrastive learning with counterfactual explanations for radiology report generation. In *European Conference on Computer Vision*, pages 162–180. Springer, 2024. 2

[28] Shanhao Li, Bang Yang, and Yuexian Zou. Adaptive curriculum learning for video captioning. *IEEE Access*, 10:31751–31759, 2022. 3

[29] Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal representation learning via maximization of local mutual information. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 273–283. Springer, 2021. 4, 6

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[31] Aohan Liu, Yuchen Guo, Jun-hai Yong, and Feng Xu. Multi-grained radiology report generation with sentence-level image-language contrastive learning. *IEEE Transactions on Medical Imaging*, 43(7):2657–2669, 2024. 2

[32] Fenglin Liu, Shen Ge, Yuexian Zou, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. *arXiv preprint arXiv:2206.14579*, 2022. 3

[33] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019. 6

[34] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10348–10359, 2025. 2

[35] Chongwen Lyu, Chengjian Qiu, Kai Han, Saisai Li, Victor S Sheng, Huan Rong, Yuqing Song, Yi Liu, and Zhe Liu. Automatic medical report generation combining contrastive learning and feature difference. *Knowledge-Based Systems*, 305: 112630, 2024. 2

[36] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40, 2022. 2

[37] Hoang Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated generation of accurate & fluent medical x-ray reports. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3552–3569, 2021. 2

[38] Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. Factual accuracy is not enough: Planning consistent description order for radiology report generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7123–7138, 2022. 2

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[40] Gustavo Penha and Claudia Hauff. Curriculum learning strategies for ir: An empirical study on conversation response ranking. In *European conference on information retrieval*, pages 699–713. Springer, 2020. 7

[41] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 1162–1172, 2019. 7

[42] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*, 2019. 5

[43] Gabriel Reale-Nosei, Elvira Amador-Domínguez, and Emilio Serrano. From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. *Medical Image Analysis*, page 103264, 2024. 2

[44] Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 2024. 2

[45] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022. 3

[46] Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 643–655, 2025. 2

[47] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 2, 6, 7

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[49] Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision*, pages 563–579. Springer, 2022. 2

[50] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021. 3

[51] Xinyi Wang, Grazziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. A survey of deep-learning-based radiology report generation using multimodal inputs. *Medical Image Analysis*, page 103627, 2025. 2

[52] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2433–2442, 2021. 2

[53] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023. 2

[54] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021. 7

[55] Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 2023. 3

[56] Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8664–8672, 2025. 2

[57] Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11595–11603, 2022. 3

[58] Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. Writing by memorizing: Hierarchical retrieval-based medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5000–5009, 2021. 2

[59] Xinyu Yang, Tilo Burghardt, and Majid Mirmehdi. Dynamic curriculum learning for great ape detection in the wild. *International Journal of Computer Vision*, 131(5):1163–1181, 2023. 3

[60] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023. 6

[61] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12910–12917, 2020. 2

[62] Qin Zhou, Guoyan Liang, Xindi Li, Jingyuan Chen, Zhe Wang, Chang Yao, and Sai Wu. Learnable retrieval enhanced visual-text alignment and fusion for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22529–22538, 2025. 2